mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections

Year: 2022 | Citations: 264 | Authors: Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan

---

Abstract

Large-scale pre-trained foundation models have been an emerging paradigm for building artificial intelligence (AI) systems, which can be quickly adapted to a wide range of downstream tasks. This paper presents mPLUG, a new vision-language foundation model for both cross-modal understanding and generation. Most existing pre-trained models suffer from inefficiency and linguistic signal overwhelmed by long visual sequences in cross-modal alignment. To address both problems, mPLUG introduces an effective and efficient vision-language architecture with novel cross-modal skip-connections.mPLUG is pre-trained end-to-end on large-scale image-text pairs with both discriminative and generative objectives. It achieves state-of-the-art results on a wide range of vision-language downstream tasks, including image captioning, image-text retrieval, visual grounding and visual question answering. mPLUG also demonstrates strong zero-shot transferability on vision-language and video-language tasks. The code and pre-trained models are available at https://github.com/alibaba/AliceMind